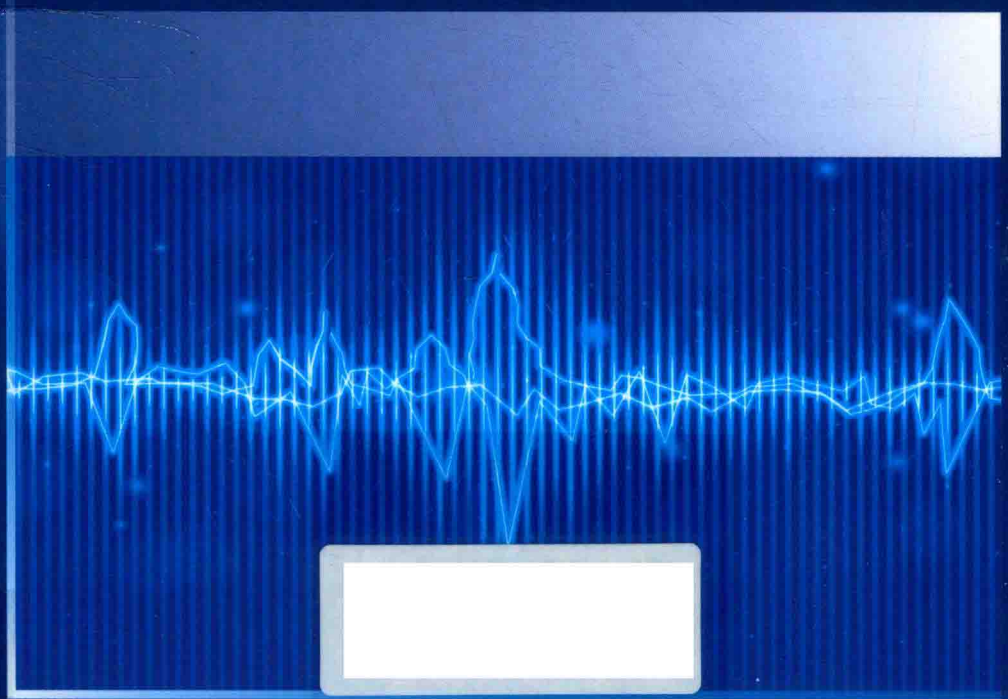
语音识别环境失配补偿技术

何勇军　著



科学出版社

语音识别环境失配补偿技术

何勇军　著

科学出版社

北京

内 容 简 介

本书系统论述了语音识别中的环境失配问题和补偿方法。全书共 8 章，内容包括环境失配问题研究的内容和意义以及研究现状；基于隐马尔可夫模型的语音识别；基于形态成分分析的鲁棒语音活动检测；基于稀疏编码的加性噪声补偿；基于语音字典的评价与优化的补偿方法；信道畸变两个子问题的划分与补偿；高斯依赖的信道畸变补偿；基于对数运算线性分段函数的联合补偿。

本书可作为高等院校从事人工智能和语音信号处理相关研究的硕士、博士研究生的参考书，也可供从事计算机信息科学、人工智能和数据挖掘的科技人员和工程人员参考。

图书在版编目（CIP）数据

语音识别环境失配补偿技术 / 何勇军著. — 北京：科学出版社，2019.11

ISBN 978-7-03-057543-2

Ⅰ. ①语… Ⅱ. ①何… Ⅲ. ①语音识别－研究 Ⅳ. ①H012

中国版本图书馆 CIP 数据核字(2018)第 100526 号

责任编辑：王 哲 / 责任校对：彭珍珍

责任印制：吴兆东 / 封面设计：迷底书装

科学出版社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2019 年 11 月第 一 版 开本：720×1 000 1/16

2020 年 1 月第二次印刷 印张：11 插页：1

字数：210 000

定价：109.00 元

（如有印装质量问题，我社负责调换）

前　言

语音信号处理是人工智能领域的一个重要分支，主要研究识别语音信号的内容，在智能人机接口、机器人语音理解、语音内容分析、军事侦察、工业控制、听写机、语言辅助学习和呼叫中心等领域有着广泛应用。尤其是近年来，随着深度学习的迅猛发展，语音识别已经走出了实验室，并广泛应用于各个领域。然而，语音不可避免地会受到噪声、信道畸变和各种编码差异的影响，导致训练和应用的环境失配，这将降低语音识别系统的识别率，因此迫切需要用信号增强和信道畸变补偿技术来提高语音识别系统的环境鲁棒性。

环境失配补偿一直是语音识别领域的一个研究热点，吸引了大量科研人员参与研究。随着语音识别技术的应用推广，越来越多的研究人员和工程技术人员开始从事语音识别相关的研究和开发工作。

从早期的研究视角来看，语音识别环境鲁棒性问题是由噪声、信道畸变、编码差异等因素引起的。随着研究的深入，人们发现这归根结底是训练环境和测试环境不一致的问题。在训练阶段，不可能把所有可能环境下的数据都采集到，只能获得一些典型环境下的数据。而在识别阶段遇到的环境是多样的。这种不一致将导致语音识别系统的性能急剧降低。其结果是一个语音识别系统在实验环境下能取得很好的效果，一旦到了现实应用环境，其错误率不断上升。本书的首要任务是要阐述清楚环境失配问题的实质。

语音活动检测是语音的预处理技术之一，对提高语音识别的效率和准确率有着重要作用。在干净语音情况下的语音活动检测并不是一个难题，但当存在噪声时，语音活动检测就变得困难。语音识别鲁棒性问题首先要解决的是噪声环境下的语音活动检测问题。本书首先立足于这一问题，给出了基于稀疏编码的语音活动检测方法。

噪声是引起环境失配最常见的因素。因此，噪声的补偿也是本书要探讨的一个重要问题。对于噪声的补偿，一个很自然的想法就是去噪。目前科研人员在语音去噪方面提出了许多方法。这些方法在一定程度上能降低噪声，提高信噪比，其评价有主观和客观两个方面。主观评价主要通过人对去噪后的语音打分，着眼于语音的可懂度和听觉感受。客观评价着眼于信噪比的提高。这两种评价指标都不是直接以提高语音识别系统的准确率为目的，导致大多数语音增强方法难以在语音识别的噪声补偿上取得好的效果。稀疏编码既可用于信号域的语音增强，也可作为分类器来使用，使语音识别鲁棒性得到了较大的提升。本书对稀疏编码在噪声补偿方面的应

用进行了详细阐述，可为相关研究和工程人员提供借鉴。

信道畸变是不可避免的。不同的通信信道、不同的麦克风以及不同的编码策略等，都会引起信道畸变。这一问题在电话语音的识别上比较突出。信道的影响可以看作一个滤波器。不同的信道有不同的频率响应曲线。典型的窄带电话信道的截止频率为 200～3400Hz，而宽带信道的上限截止频率超过 8000Hz。而且各种信道对相同子带的增益也是不同的。此外，不同的编码方式解码后的语音在频谱上也是不同的。这些不同都会造成信道畸变，这也是本书要解决的问题之一。

全书共 8 章。第 1 章是绪论，主要介绍研究的目的和意义，以及本领域研究的进展情况和存在的问题。第 2 章是基于隐马尔可夫模型的语音识别，主要介绍语音识别基本理论和方法，以及环境失配的数学描述。第 3 章是基于形态成分分析的鲁棒语音活动检测，给出一种基于稀疏编码的鲁棒语音活动检测方法。第 4 章是基于稀疏编码的加性噪声补偿，主要涉及采用稀疏编码实现加性噪声的补偿，提高语音识别的鲁棒性。第 5 章是基于语音字典的评价与优化的补偿方法，提出一系列语音字典的评价指标，并在此基础上给出了一个系统的字典评价方法，提出采用提高评价指标的方法对字典做进一步优化。第 6 章是信道畸变两个子问题的划分与补偿，主要介绍信道畸变的补偿，通过划分两个子问题并分别进行处理，达到补偿目的。第 7 章是高斯依赖的信道畸变补偿，给出一种高斯依赖的信道畸变补偿方法，对信道畸变进行更精细化的补偿。第 8 章是基于对数运算线性分段函数的联合补偿，给出一种在对数运算情况下对非线性畸变模型分段线性化的补偿方法，所提出的方法既有信号、特征域的补偿，也有模型域的补偿。

本书的撰写是由作者及其团队集体共同努力完成的。需要特别感谢的是研究生梁隆恺、赵晶、卢玉、余莲、卢祎、张雪媛和郭云雪等。同时感谢谢怡宁老师和黄金杰老师的大力支持。

由于作者水平有限，书中难免存在不妥之处，希望广大读者批评指正。

何勇军

2019 年 8 月

目　录

第1章 绪 论

1.1 研究的内容和意义

随着信息技术的迅猛发展，信息的获取、交互与处理已成为当今社会发展的一大动力。以计算机为中心的信息技术不断地改变着人们的生活方式，这一过程被誉为信息革命，并已成为继工业时代后的知识时代里的一大里程碑[1]。语音是人类最自然、最常用的信息交流方式。无论是在生活中还是在互联网上，语音作为主要媒体之一，承载着大量的有用信息。因此，对语音中的信息进行分析、处理和识别无疑具有广阔的应用前景。作为语音处理的支撑技术之一，语音识别以识别语音信号并将其转换成文字为目标，在智能人机接口、机器人语音理解、语音内容分析、军事侦察、工业控制、听写机、语言辅助学习和呼叫中心等领域有着广泛应用。

语音识别的历史可以追溯到20世纪30年代初，当时的研究者们尝试识别特定的声音，并开始从声学角度识别音素或数字等，其任务还局限于小词表孤立词识别。20世纪60～80年代，语音识别技术快速发展，典型的进展是基于线性预测的频谱分析[2,3]、基于线性规划的语音时间对齐方法[4]以及矢量量化的成功应用[5]，识别任务也发展到了中等规模的孤立词识别和连接词识别[6]。20世纪最后20年里，语音识别技术取得了长足进步，最重要的里程碑是隐马尔可夫模型(hidden Markov model，HMM)在语音识别领域的成功应用[7,8]，辅以前向后向算法、K均值训练算法、维特比解码算法、基于神经网络的条件概率估计方法[9]和各种模型自适应方法的提出，使语音识别迈向了非特定人大词表连续语音识别(large-vocabulary continuous speech recognition，LVCSR)的新阶段。近十年来，语音识别进一步飞速发展，出现了区分性训练、不确定性解码、噪声鲁棒性以及机器学习等新技术，使语音识别走出了实验室并逐步走向现实应用。

在这一背景下，世界各国为了抢占语音识别领域的制高点，纷纷投入了大量人力物力支持语音识别的相关研究和产品开发，并提出了与之相关的若干重大发展计划，例如，欧洲的CHIL(computers in the human interaction loop)和AMI(augmented multi-party interaction)计划、美国的VACE(video analysis and content extraction)和CALO(cognitive assistant that learns and organizes)计划等[10]。美国国防部高级计划研究署和国家安全局也支持了一大批语音识别相关的研究。相应的，美国标准技术研究署(National Institute of Standards and Technology, NIST)在21世纪初陆续举办了一

系列语音识别相关的评测[11-13]，其任务从最初的朗读语音到广播语音，再到后来的交谈式电话语音，然后发展到了目前真实场景下的会议语音。结果表明，虽然语音识别取得了巨大进展，但在特征表示、声学模型以及环境鲁棒性等方面仍然存在许多问题有待解决。此外，世界各国的著名研究机构和公司企业纷纷向语音识别系统的研发投以巨资，研制出了各种大词表连续语音识别系统，如卡内基梅隆大学的 Sphinx 系统，剑桥大学的 HTK、OG 系统和 DARGON 系统，IBM 的 ViaVoice 系统，微软的 Whisper 和 Office 语音录入系统等。特别是近期苹果公司集成在 iPhone 上的语音助手 Siri、谷歌的语音搜索系统和科大讯飞发布的语音云端系统等，已成为智能人机交互的典范，进一步刺激了语音识别相关技术的发展。

我国在国家自然科学基金、863 计划和 973 计划中也支持了一批与语音识别相关的研究工作，有力地促进了国内语音识别技术的发展。在这些项目的支持下，国内的中国科学院、清华大学[14]、北京大学、中国科学技术大学[15]、哈尔滨工业大学[16]、北京邮电大学、华南理工大学[17]、西北工业大学、解放军信息工程大学[18]等也纷纷开展可卓有成效的研究。从 2008 年起，国家自然科学基金委紧跟国际学术前沿，适时启动了“视听觉信息的认知计算”重大研究计划，以推动视听觉认知相关的计算模型和计算方法的研究。该计划的一个重要方向就是语音的感知和理解，集中体现了国家对语音识别研究的重视程度。

经过数十年的发展，语音识别技术取得了巨大成就。在理想环境下，目前的小词表以及中等词表语音识别系统的识别率能达到 99%以上，LVCSR 系统识别率也能超过 95%[19]。但是，在训练条件和测试条件不匹配时，系统识别率将急剧下降。造成环境不匹配的因素众多，典型的有声学环境失配、说话方式差异、说话人差异、词汇量和领域差异。其中，声学环境失配是导致系统性能下降的主要原因，也是语音识别系统走向应用所面临的最大挑战。声学环境是指语音从产生到成为待识别数字信号这一转变过程[20]。如图 1-1 所示[21]，语音可能受说话人的状态比如疲劳、压力、生病等因素影响而发生改变；当外界存在噪声时，说话人还会自然地提高音量，即产生所谓的 Lombard 效应[22]；发出的语音与噪声同时进入麦克风使语音受到污染；各种麦克风的传输性能不一样，也会对语音造成影响；语音可能通过信道传输，

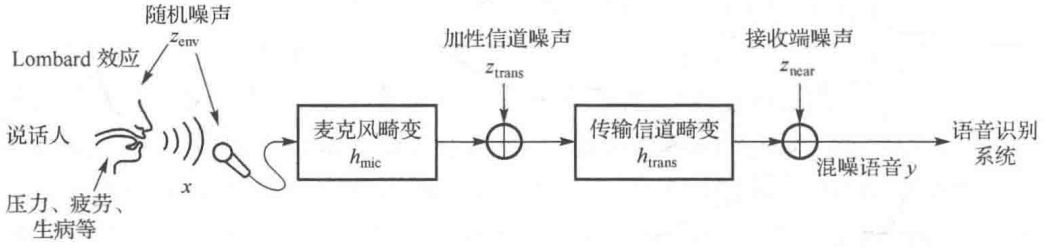


图 1-1　声学环境

在信道中也存在加性噪声，且不同的传输信道也会对语音产生不同的影响；在接收端还可能再次受加性噪声的干扰。考虑上述因素，语音受到的畸变可表示为

$$y(\tau)=((G(x(\tau))+z_{\text{env}}(\tau))*h_{\text{mic}}(\tau)+z_{\text{trans}}(\tau))*h_{\text{trans}}(\tau)+z_{\text{near}}(\tau) \tag{1-1}$$

式中，$*$是卷积运算，τ是时间序号，$y(\tau)$是畸变后的语音，$x(\tau)$是干净语音，$G(\cdot)$是说话人状态影响算子，$z_{\text{env}}(\tau)$是说话环境随机噪声，$h_{\text{mic}}(\tau)$是麦克风的冲击响应，$z_{\text{trans}}(\tau)$是传输信道的加性噪声，$h_{\text{trans}}(\tau)$是传输信道的冲击响应，$z_{\text{near}}(\tau)$是接收端的加性噪声。本章将不涉及说话人状态受到影响的情况，因此，$G(\cdot)$的作用不予考虑。通过将多个加性噪声合并成一个加性噪声$v(\tau)$，将多个卷积噪声合并成一个线性信道噪声$h(\tau)$，式(1-1)可简化为目前鲁棒语音识别领域普遍采用的畸变模型[23,24]

$$y(\tau)=h(\tau)*x(\tau)+v(\tau) \tag{1-2}$$

在语音识别中，首先需要提取待识别语音信号的特征参数，然后在声学模型上进行匹配，最后得到识别结果。当待识别语音与训练语音处于同一声学环境，则称环境匹配，否则称为环境失配。同一语音在不同环境中其信号、特征和概率分布都会发生改变[25]。如图1-2所示，虚线上方为干净环境α，下方为噪声环境β。环境失配相当于在信号域、特征域和模型域存在畸变函数$D_1(\cdot)$、$D_2(\cdot)$和$D_3(\cdot)$分别改变了信号、特征或模型。在图1-2中，用模型Λ_α识别F_α或用Λ_β识别F_β时称为匹配识别；用Λ_α识别F_β时称为失配识别。从根本上说，环境失配改变了语音特征的概率分布，使其与原有声学模型失配，因而导致识别系统的性能下降。从式(1-2)可以看出，造成环境失配的主要原因在于噪声的存在，这种噪声可能是加性的$v(\tau)$，可能是卷积性的$h(\tau)$，也可能是这两者的混合[26]。诸如训练环境没有噪声而测试环境存在噪声，或者训练环境和测试环境存在不同的噪声，都会引起环境失配。

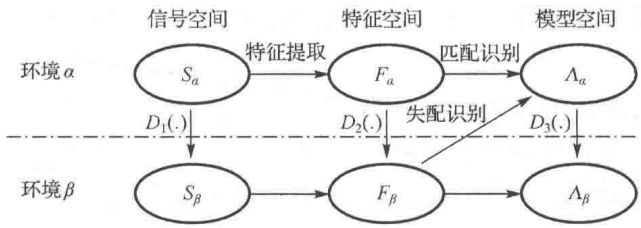


图1-2 训练和识别环境失配

环境失配必然导致语音特征参数的分布存在偏差进而影响系统性能。研究表明，当存在信噪比为10dB的加性噪声时，不做任何补偿的孤立词识别系统的识别率从99%降到49%[27]；随着信噪比继续降低，语音识别系统将无法使用。现实应用环境复杂多变，环境失配不可避免，只有提高系统的环境鲁棒性才能使语音识别技术真正走向应用。

相比而言，人的听觉在噪声情况下具有很强的鲁棒性。研究发现，虽然在噪声

环境下人的误识率也会增加，但其降低的速度要大大慢于当前的语音识别系统[28]。大量研究者也尝试模仿人的听觉机理来提高系统鲁棒性，虽取得了一定的进展，但人的听觉机理和处理信息的方式至今尚未被深刻理解。因此，要使机器语音识别具有类似人类听觉系统的鲁棒性仍需要大量有效而深入的研究。本书正是针对语音识别中的环境失配问题，研究并提出增强系统环境鲁棒性的有效方法，从而推动语音识别技术走向现实应用。

1.2　国内外研究现状与分析

式(1-1)是声学环境的数学描述，全面刻画了加性噪声和信道畸变。在现实应用中，环境失配存在三种可能的情况，即加性噪声、信道畸变和二者同时并存。相应的，目前的环境失配补偿方法也可以大致分为加性噪声补偿、信道畸变补偿和联合补偿三类。下面将按照这一分类方式阐述环境失配补偿的研究现状。

1.2.1　加性噪声补偿方法

加性噪声在时域和频域都呈加性，但在对数频域和倒谱域因对数运算呈高度非线性。加性噪声的补偿大致可以分为特征规正、特征增强和模型补偿三种类型。特征规正试图将语音特征规正到受噪声影响较小的特征空间。比如，倒谱均值规正(cepstral mean normalization, CMN)[29,30]通过在一个时间段内统计倒谱均值，然后从各帧中减去这一均值以减小噪声对特征的影响。倒谱均值方差规正(cepstral mean variance normalization, CMVN)[31]同时规正倒谱特征的均值和方差，使得规正后的特征均值为0，方差为1。倒谱直方图规正(cepstral histogram normalization, CHN)[32]则用倒谱直方图代替其概率分布，通过规正直方图到已知的形状以达到规正倒谱概率分布的目的。

特征增强则试图从混噪语音中估计干净语音特征，然后用于匹配识别。这类方法要么先增强信号再提取特征，要么直接从混噪语音特征中估计干净语音特征。在信号增强方面，谱减[33]和非线性谱减[34]是出现较早且研究较深的语音增强方法。谱减法首先用噪声段估计噪声谱，然后从各帧语音谱中减去噪声谱。这种方法虽然可以提高信噪比，但不可避免地会形成残留噪声(音乐噪声)。非线性谱减通过比较噪声谱和混噪语音谱后采取分段处理的策略，能有效减少残留噪声。但是，上述两种方法皆不适合快速时变的噪声环境。维纳滤波在满足干净语音和输出语音的均方误差最小条件下，设计线性滤波器处理混噪语音，被广泛用于语音增强。这一方法已被成功用于欧洲分布式语音识别编码与传输标准[35]。自适应滤波和卡尔曼滤波[36]对时变噪声表现出了一定的跟踪能力。子空间降噪方法[37]将混噪语音投影到一个子空间，并认为干净语音和噪声分布在不同的分量上，通过压制噪声所代表的分量实

现降噪。如果硬件和应用场合允许，基于麦克风阵列的语音增强也可用于信号去噪并提取特征。麦克风阵列可以首先确定声源位置，然后采用空域滤波的方式，通过自适应滤波器增强某一位置的声音同时抑制其他位置的声音，典型的方法有基于波束形成的方法[38]、多通道维纳滤波[39]、多通道子空间方法[40]和空间-时间预测法[41]等。但基于麦克风阵列的增强方法对硬件有较高要求，也不适合处理已存在的单通道语音信号。语音信号在时域或频域被增强后，噪声成分有所减少，理论上在此基础上提取的特征受到的噪声影响也会变小。

在特征增强方面，基于数据驱动的方法因其效果明显而受到广泛关注。这类方法在学习干净语音和混噪语音的概率分布时，需要干净环境和噪声环境下同时录制并对齐的立体声数据。基于码字的倒谱规正方法(NR-dependent cepstral normalization, SDCN)[23]假定干净语音的特征服从高斯混合模型(Gaussian mixture model, GMM)分布，通过迭代的方式在混噪语音上估计噪声参数，然后补偿混噪特征。SPLICE[42]通过学习混噪语音和干净语音特征的联合分布，然后用混噪特征得到干净特征的最小均方误差(minimum mean squared error, MMSE)估计。另一种使用立体声数据的补偿方法是 RATZ[43]，该方法仍然用高斯混合函数为干净语音建模，并使用 MMSE 估计获得干净语音的估计值。尽管基于立体声数据的补偿方法能取得较好效果，但需要预先知道噪声环境并为此准备带标注的立体声数据，应用条件苛刻，无法适用于未知声学环境。

目前的多数方法立足于对噪声的建模与估计，然而，噪声的时变特性及其与语音之间的复杂作用方式使得对噪声的建模与估计非常困难且不可靠。丢失数据技术[44]不对噪声做任何假设，也无须为噪声建模与估计，而只需要知道噪声对语音频谱污染的严重程度即可实现补偿。该方法在前端用一个标记向量将语音特征分为可信的和不可信的两部分，在后端要么丢弃不可信部分，直接利用可信部分进行识别，要么利用统计的方法重估丢失部分特征，然后利用可信部分和重估部分进行识别[44,45]。标记向量根据语音谱的局部信噪比确定，最初只含有元素 0 和 1，分别表示对应特征完全丢失和完全可信，而后扩展到可以取 0 到 1 之间的任何数，即所谓的软决策[46,47]。丢失数据方法最初运行在对数频谱域，后来又被扩展到了倒谱域[48,49]。在丢失频谱的重构方面，可以采用基于最大后验概率或基于聚类的估计方法[45]。在估计过程中，各种先验知识也可融入其中，最常用的是时频相关性。比如文献[50]在频谱上组合传统的基于频率相关性和基于时间相关性的方法重构丢失特征，并对二者做一个合理的折中，取得了一定的效果。基于丢失数据技术的方法虽然不需要对噪声进行建模与估计，在信号处理上也类似于人耳对声音的处理方式，但在计算标记向量时需要判断混噪频谱是否可信，这同样是一个难以解决的问题。与丢失数据方法类似的另一类方法是不确定解码[51]。该方法根据信噪比对不同的特征赋予不同的置信度，在后端充分考虑这些置信度以达到提高系统鲁棒性的目的。

近年来，随着系数分解与重构理论的成熟，稀疏编码(sparse coding)[52-54]在信号处理的各个领域表现出了巨大潜力。该技术在稀疏性准则下将信号用一组基元信号线性表示，获得信号的稀疏表示(sparse representation)。其中，每个基元信号称为一个原子(atom)，所有原子组成的集合称为字典(dictionary)。稀疏性是指信号被分解到某个字典上时，仅有少量原子的系数不为零。现实中的大量信号，如语音、图像等都满足或近似满足稀疏性[55]；研究表明，人的感知神经系统总是从海量神经元中仅激活极少一部分[56,57]以实现对外部刺激的编码。这意味着人的感知系统在处理信号时也利用了稀疏性原则。稀疏性似乎是信号本身具有的特点，是一种先验知识；而稀疏编码的有效性正是因为利用了这种先验知识。作为一种新兴的技术，稀疏编码可广泛用于信号压缩、分析、去噪和分离等[58]。尤其是近年来，该技术在图像处理及模式识别领域的成功应用[59]极大地增强了研究者们深入研究的信心。

用稀疏编码增强语音频谱需要解决三个问题，即字典构建、稀疏分解和频谱重构。字典构建是稀疏编码的首要问题，目的在于选取有代表性的基元信号(原子)构成字典。目前的方法大致可分为基于选择的和基于学习的两类。基于选择的方法从预先定义的基函数中直接选取需要的原子组成字典[60-62]。基于学习的方法则在满足重构误差要求的情况下，从大量数据中学习一组能稀疏表示信号的原子组成字典[63,64]。稀疏分解的目的在于将信号表示为各原子的线性组合，其求解过程是非线性的；典型的方法有正交匹配追踪算法(orthogonal matching pursuit, OMP)[65]、匹配追踪算法(matching pursuit, MP)[52]和基追踪去噪算法(basis pursuit de-noising, BPDN)[53,66]等。谱重构则利用稀疏表示和字典重构干净频谱，其过程是线性的。近年来，在语音及其特征增强方面，稀疏编码开始被用于语音增强[67,68]和鲁棒语音识别[69,70]。但在用法上大多是将稀疏编码作为工具简单使用，缺乏针对语音信号特殊性的有效分析和应用。

1.2.2　信道畸变补偿方法

当加性噪声可被忽略时，信道畸变在时域表现出卷积性，而在频域则表现出乘积性。因此，目前的方法普遍认为，信道畸变在对数频谱域或在梅尔倒谱域是一个加性的常量。其补偿方法也可分为特征规正、特征补偿和模型补偿。特征规正类方法中最典型的是 CMN 和相关谱滤波(relative spectra, RASTA)[71]。CMN 认为信道影响存在于特征的直流分量中，从各帧中减去均值即可消除直流分量，从而去除信道影响。RASTA 认为信道产生的畸变存在于信号的慢变分量中，通过设计一个低通滤波器抑制信号的慢变分量即可抑制信道畸变。CMN 和 RASTA 是两种标准的规正处理方法，被广泛用于语音识别系统中，但文献中大量实验表明 RASTA 在没有畸变时反而会降低系统识别率。此外，文献[72]提出通过一个梅尔频域的带通滤波器逐帧处理畸变语音。该方法首先通过区分性函数做重要性分析求得滤波参数，然后利

用传统的特征轨迹滤波方法去除信道影响。实验表明该方法与 CMN 和 RASTA 组合后还能进一步提高识别率。

在特征补偿方面，基于贝叶斯框架的信道参数估计方法[73]假定信道畸变为一个加性常量，并分别在假定语音服从高斯、高斯混合模型和隐马尔可夫模型的情况下，用最大似然、最大后验概率从畸变语音中估计信道偏移量。补偿时从特征中减去信道偏移量，达到补偿信道畸变的目的。在模型域，信号偏移量去除法(signal bias removal，SBR)[74]在声学模型上用期望最大化算法(expectation maximization, EM)迭代地估计信道偏移量，然后在模型域修改高斯均值，最后用修改后的模型识别畸变语音。文献[75]用丢失数据技术，将倒谱特征的静态和动态参数表示成对数频谱的线性组合。该方法基于丢失数据框架，利用噪声谱的可信部分，在对数频谱域或倒谱域用最大似然准则估计一个直流偏移量以实现信道畸变的补偿。

近年来，研究者们开始关注语音识别中另一类由信道引起的畸变，即窄带语音畸变问题[76-80]。当待识别语音的带宽比训练语音的带宽窄时，与训练语音相比，待识别语音丢失了部分频段，也将导致环境失配。这一问题在复杂信道环境(比如互联网、分布式语音识别等)中广泛存在。在增强语音质量方面，解决这一问题的方法是人工频带扩展(artificial bandwidth extension, ABE)，即将窄带语音的频带扩展，使其成为宽带语音。典型的任务就是将带宽为 0～3.4kHz 的窄带语音扩展为 0～8.0kHz 的宽带语音。ABE 需要解决的问题是利用现有的窄带数据，重构 3.4～8.0kHz 频段的频谱。目前大多数方法都基于语音的源-滤波器产生模型。这类方法先产生一个激励信号，然后用一个滤波器模拟声道处理激励信号以生成高频段频谱，最后组合窄带信号和重建的高频信号，获得宽带信号。在激励信号方面，目前常用的方法有频谱折叠、频谱变换和非线性处理等[81]，也可使用正弦合成[82]或噪声调节模拟激励源[83]。声道通常用一个全极点滤波器模拟，而滤波器参数可以是线性频谱或倒谱系数[82]。窄带信号是已知的，需要用窄带特征估计高频段滤波参数。常用的有基于码本[82,84,85]、高斯混合模型[86,87]、隐马尔可夫模型[88,89]以及人工神经网络[82,90]等方法。

尽管从信号增强的角度已经有大量工作扩展频带，但研究表明，直接用上述方法扩展频带后提取的特征在提高系统识别率上非常有限[91]。其原因在于用 ABE 方法重构的高频段虽然能提高人的主观听觉，但这与提高识别准确率的目标不一致，提取的特征在识别时仍然存在失配问题。研究者们尝试直接以语音识别为目的来补偿带宽失配。文献[77]采用特征域的限定最大似然线性回归(constraint maximum likelihood linear regression, CMLLR)将宽带特征直接转变为窄带特征，然后用于窄带声学模型的训练。文献[76]将基于 GMM 的频段扩展方法和基于 HMM 的声学模型结合在一起，实现宽带声学模型对窄带语音的识别。文献[78]和文献[79]通过训练数据学习一组矫正函数以实现畸变语音特征向干净语音特征的变换；而文献[80]则在

丢失数据框架下利用语音频谱的时频相关性，实现丢失频段的重构。尽管上述方法取得了一定的效果，但其本质上是特征域补偿方法，而且严重依赖于训练数据。一方面需要大量训练数据，另一方面要求语音带宽稳定。此外，这些方法只是单纯地补偿窄带引起的畸变，而未能同时考虑常规信道畸变，导致这些方法仅适用于单一的带宽失配场合，不适合各种畸变同时存在的复杂环境。

1.2.3　联合补偿方法

当信道畸变和加性噪声同时存在时，需要对这二者同时进行补偿。在特征域可以选用能有效表示信号的特征；在模型域可以用数据驱动的方式进行各种自适应，也可以利用畸变模型估计噪声参数然后补偿声学模型。

鲁棒特征提取旨在有效表示信号并提取受噪声影响较小的特征以提高系统鲁棒性。常用的特征有基于人耳听觉特性的梅尔频率倒谱系数(Mel frequency cepstrum coefficient, MFCC)[92]、基于自回归模型的线性预测系数(linear prediction coefficient, LPC)[93]和基于听觉感知的相关谱感知线性预测系数(relative spectra perceptual linear prediction, RASTA-PLP)[94]。此外，目前较新的 Teager 能量倒谱系数(Teager energy cepstral coefficient, TECC)[95]采用一种稠密平滑的滤波器组以及可变的能量计算策略，取得了比 MFCC 更强的噪声鲁棒性。而瓶颈特征(bottleneck feature, BF)[96,97]则采用含有少量显层和大量隐层节点的瓶颈状神经网络生成新特征，虽然计算复杂度较高，而且需要训练神经网络，但文献表明，这类特征具有较强的鲁棒性。

虽然研究者们在特征提取方面做了大量尝试，提出了一系列特征，也取得了一定的效果。但目前仍然不清楚哪些参数携带着最有用、最具鲁棒性的信息。因为特征参数能表示语音，在一定程度上也能表示噪声，换言之，特征参数中不可避免地混有噪声，引起识别率的下降。

由于噪声的影响在倒谱域呈现高度非线性，Moreno[98]等提出用矢量泰勒级数(vector Taylor series, VTS)将非线性关系近似展开成线性，同时假定干净语音服从高斯混合分布，并用每句语音在线估计信道参数和噪声参数然后补偿特征。该方法无需额外训练数据，能在短时内动态补偿噪声，具有一定的优势。后来，VTS 这一思路被广泛用于模型域补偿。

特征域补偿方法试图补偿语音特征使其与模型相匹配，而模型域方法则修改声学模型使其适应输入的特征。匹配训练的思路最为直观，即直接采集某一噪声环境下的语音进行标注然后重新训练模型，用来识别这一噪声环境下的语音。这意味着每遇到新的环境，匹配训练都需要重新采集并标注数据，训练模型。这样做能保证识别率，但耗时耗力，无法推广。多重风格训练[99]采集所有可能的噪声环境下的数据训练声学模型。这种方法也有明显弊端：一方面，未知噪声环境的噪声类型无法穷尽，另一方面，过多的数据会导致声学模型的区分能力降低。匹配训练和多重风